Syracuse University, School of Information Studies

M.S. Applied Data Science

Portfolio Milestone

Bryan J. Crigger

SUID: 255676562

Bryan.crigger7@gmail.com

Contents

[Introduction 3](#_Toc161688073)

[Applied Machine Learning 4](#_Toc161688074)

[Project Description 4](#_Toc161688075)

[Learning Objectives & Outcomes: 5](#_Toc161688076)

[Natural Language Processing 5](#_Toc161688077)

[Project Description 5](#_Toc161688078)

[Learning Objectives & Outcomes: 6](#_Toc161688079)

[Big Data Analytics 7](#_Toc161688080)

[Project Description 7](#_Toc161688081)

[Learning Objectives & Outcomes: 8](#_Toc161688082)

[Deep Learning in Practice 9](#_Toc161688083)

[Project Description 9](#_Toc161688084)

[Learning Objectives & Outcomes: 10](#_Toc161688085)

[Conclusion 11](#_Toc161688086)

[Appendix 11](#_Toc161688087)

# Introduction

The M.S. in Applied Data Science program through Syracuse University’s School of Information Studies is designed to not only provide the theoretical foundation of Data Science concepts, but also strongly focuses on the application of data science on real-world problems. The purpose of the portfolio is to showcase how these principles and the additional program learning goals were applied through my course work. The projects I have included in my portfolio reflect my ability to apply Data Science techniques and methods towards a wide range of domains. The projects chosen are from the classes Applied Machine Learning, Natural Language Processing, Big Data Analytics, and Deep Learning in Practice. Each project section includes the summary of what the project entailed and how the project aligned with the program’s learning objectives. The learning objectives for the Applied Data Science Program were:

1. Collect, store, and access data by identifying and leveraging applicable technologies.
2. Create actionable insight across a range of contexts, using data and the full data science life cycle.
3. Apply visualization and predictive models to help generate actionable insights.
4. Use programming languages such as R and Python to support the generation of actionable insight.
5. Communicate insights gained via visualization and analytics to a broad range of audiences.
6. Apply ethics in the development, use and evaluation of data and predictive models.

# Applied Machine Learning

## Project Description

The project for the Applied Machine Learning class focused on analyzing opioid data from the Centers for Medicare & Medicaid Services (CMS), which was accessed through cms.gov and Kaggle. The opioid crisis in the United States continues to be a significant public health issue. Despite widespread awareness of the dangers of opioids, overdose deaths have quadrupled since 1999. To address these issues, the analysis aimed to discover patterns in opioid prescriptions, focusing on prescriber types, what specific opioids were most often prescribed, and to identify high-risk areas and prescriber types. The original dataset used consisted of over 23 million records from 2013. However, the analysis mainly focused on prescriber data from 2019, comprising of 971,968 observations and 254 variables, including prescriber information and details on 250 common opioid and non-opioid drugs.

The analysis employed several data mining techniques like Association Rule Mining, Clustering, and Classification Algorithms. The initial findings revealed a highly skewed distribution of opioid prescriptions, with a small percentage of prescribers responsible for a large number of prescriptions, particularly in the Southern United States. KMeans Clustering and Naive Bayes modeling were both used to group prescribers based on their prescribing behaviors and to identify conditional probabilities associated with opioid prescribing. The project highlighted that clinicians with the specialties of family practices, internal medicine, physician assistants, orthopedic surgeons, and general surgery clinics were the most significant opioid prescribers. The findings emphasized the need for careful monitoring of opioid prescription amounts; increased access to rehabilitation, especially in high-prescriber states like Florida, Texas, California, and North Carolina; and acknowledgment that socio-economic factors like level of education and race have on opioid prescription and overdose rates.

## Learning Objectives & Outcomes:

This project on opioid prescribing patterns in the United States serves as a prefect case of how data science can be applied to address complex societal issues. A key learning outcome of this project was the ability to effectively collect, clean, and produce actionable insights from a vast amount of data. The data used was from the CMS and consisted of over 23 million records, which demonstrate the ability to manage large-scale data efficiently. This aligns with the learning objective of identifying and utilizing appropriate technologies for data handling. Additionally, the full data science life cycle—from data collection to analysis and insight generation— was utilized, illustrating how data can be transformed into actionable insights in a societal context. The identification of high-risk prescriber types and regions for opioid prescriptions provided clear, actionable information that could inform public health policies and intervention programs.

The application of visualization and predictive modeling was central to generating these insights. The use of models like KMeans Clustering and Naive Bayes, along with the generation of various plots and heatmaps, facilitated the understanding of complex data patterns. Generating these visuals was beneficial in communicating these findings effectively to the rest of the class. Additionally, the project incorporated important ethical components, highlighting the prescribers and areas that have high levels of opioid prescriptions.

# Natural Language Processing

## Project Description

This project involved analyzing a dataset from the Enron public email corpus, containing both legitimate ("ham") and spam emails, to develop an effective classifier to accurately classify if an email is real or spam. Additional spam emails were added to the corpus to provide sufficient examples and better balance the data for training. Utilizing Python and various libraries like pandas, NumPy, nltk, and sklearn, the data was processed and cleaned, involving tokenization, filtering, and removal of stop words using NLTK’s stop word list. The project aimed to determine the most effective features and models for distinguishing between spam and ham emails.

Key findings from the project included the identification of specific words and bigrams that were indicative of spam and ham emails. For example, "pills" was frequently found in spam emails, while phrases like "investment, advice" were common in ham emails. The team explored several classifiers, including Multinomial Naive Bayes (MNB), with preprocessing steps like stemming and converting text to lowercase. The MNB classifier showed an accuracy of 87.98% but also indicated potential overfitting. Multiple vectorization techniques (term frequency, Boolean, TFIDF) and Naïve Bayes classifiers (Gaussian, Multinomial, Bernoulli) were used, finding that BernoulliNB was the most effective across all measures, regardless of the vectorizer used. K-Fold cross-validation was employed to ensure the model didn't overfit, providing a more realistic assessment of classifier effectiveness.

There were many learnings from this project. First, the preprocessing of text data, including the choice of stop word lists and techniques like tokenization and stemming, significantly improves the effectiveness of classifiers. Second, the type of classifier and vectorizer can greatly influence the accuracy and generalizability of spam detection models. Lastly, the project highlighted the importance of balancing model complexity, computational feasibility, and model explainability. These learnings highlight important data prep steps that are necessary to consider when doing NLP projects, demonstrating the complexity involved when working with text classification tasks.

## Learning Objectives & Outcomes:

Focusing on the learning objectives of creating actionable insights and effectively communicating them, the Enron email corpus project exemplified the capacity to transform complex data into meaningful information with societal and business implications. The analysis, which differentiated spam emails from legitimate emails, spelled out practical steps on how email filtering systems can be enhanced. This aspect not only showcased proficiency in extracting valuable knowledge from a large dataset but also emphasized the ability to communicate these findings clearly. By utilizing visualization techniques to represent data trends and Python for in-depth analysis, the project demonstrated a successful integration of technical acumen with the ability to convey complex insights in an accessible manner. This approach ensured that the results were not only insightful for data scientists but also able to be understood for non-technical stakeholders.

# Big Data Analytics

## Project Description

This project aimed to analyze housing data to identify factors influencing home prices, using various models like regression predictions, decision trees, and neural networks. The key recommendation was for Zillow to adopt the neural network model for predicting home prices based on specific features, emphasizing that square footage is a major contributor to home value. The project targeted providing insights to both Zillow and homebuyers, focusing on characteristics impacting house prices and thereby aiding buyers in making informed decisions.

The Zillow dataset, which we got from Kaggle, included 545 homes with details on price, area, bedrooms, bathrooms, and other features. Initial data cleaning involved formatting and outlier removal, followed by transforming categorical variables for model compatibility. Analysis revealed a strong correlation of price with area, bathrooms, and air conditioning, while factors like hot water heating and basement presence showed minimal correlation. Despite various modeling attempts, none achieved outstanding accuracy. Regression models were hindered by multicollinearity, and while random forests and decision trees provided some insights, they too fell short in model precision. The most promising results came from a neural network model using Keras, which, despite an average prediction error of 16.5%, showed the highest consistency and accuracy in predicting home values.

The project concludes that while multiple factors determine home value, no model achieved complete accuracy, highlighting the complexity of the real estate market. The recommendation for Zillow is to further develop the neural network model as a foundational tool for home value prediction. For homebuyers, understanding the impact of square footage on value is crucial. Future improvements that were considered included increasing the number of records and getting more recent housing records in the dataset for better model training, acknowledging that predicting home prices remains challenging due to dynamic market conditions and varying consumer preferences.

## Learning Objectives & Outcomes:

In this housing market analysis project, the team utilized the Kaggle platform for data collection, storage, and access, showcasing their ability to effectively manage and prepare data for analytical purposes. This proficiency in handling data was crucial in ensuring the quality and reliability of the analysis. The project was particularly successful in creating actionable insights for both Zillow and potential homebuyers, addressing key aspects of the business and societal impact of the real estate market. By employing advanced predictive models like neural networks and utilizing visualization techniques, the team was able to analyze and present factors influencing home prices in a meaningful way. Their use of Python further exemplified their technical skills in data processing and model development.

The project's strength also lay in its clear and accessible communication of insights through visualizations, making complex data understandable to a broad audience. This approach underlined the importance of effective communication in data science projects. Moreover, ethical considerations were a cornerstone of the project, with the team demonstrating awareness of potential biases and emphasizing the need for transparency in their analysis. This focus on ethics highlights the significance of responsible data usage and fairness in predictive modeling, contributing to the broader dialogue on ethical practices in data science.

# Deep Learning in Practice

## Project Description

This project for the Deep Learning in Practice class involved using Convolutional Neural Networks (CNNs) to classify images of pistachios into two main species: Kirmizi and Siirt. The aim was to leverage image recognition techniques for accurate pistachio species classification, a task critical for quality control in the food industry. Utilizing a dataset of over 2,000 images from Kaggle, the team tested multiple model variations. The best model achieved a high accuracy rate of 96.31% on test data, with an F1 score of 61% for Kirmizi and 46% for Siirt. The project employed transfer learning, a technique that leverages pre-trained models to enhance learning efficiency and accuracy. Data was organized into training, validation, and testing sets, with 70% of the data used for training and 15% each for validation and testing. Efforts were made to balance and standardize the data through oversampling and undersampling, and data augmentation was implemented using the ImageDataGenerator class from Keras.

The team experimented with various iterations of the EfficientNetB0 model to optimize it for their specific use case. This involved testing different configurations of frozen layers, dropout layers, and dense layers, as well as fine-tuning the model by unfreezing some of its layers with a low learning rate. The most successful model included a dropout layer set at 25% and a dense layer with 128 nodes using the ReLu activation function. The final fine-tuning stage involved testing different numbers of unfrozen layers to determine which configuration yielded the highest accuracy. The best results were achieved by keeping the last 15 layers frozen. The team also utilized TensorBoard for an interactive and comparative visualization of the training performance of different models. Alongside identifying the most effective model, they also assessed areas where the model underperformed and evaluated its overall performance and accuracy in classifying both types of pistachios.

## Learning Objectives & Outcomes:

In the pistachio classification project, the team effectively fulfilled key learning objectives by applying visualization and predictive modeling to generate actionable insights. Utilizing Convolutional Neural Networks, specifically the EfficientNetB0 model, they tackled the complex task of image-based pistachio species classification. The project's success in achieving high accuracy and F1 scores underscores the potent combination of sophisticated predictive models with practical applications. The team's use of TensorBoard for visualizing training performance provided clear, interpretable insights into the efficacy and accuracy of various model iterations. This approach not only enhanced the model selection process but also enabled a more transparent evaluation of the model's performance, ensuring that the chosen model was the best fit considering multiple factors.

Python's libraries, like Keras and TensorFlow, were instrumental in building, testing, and refining the CNN models. Ethical considerations were also evident in the project's execution, particularly in the handling and evaluation of data. Efforts to balance and standardize the dataset, as well as careful model tuning to avoid bias, demonstrated a commitment to fairness and accuracy for both groups of pistachio species. By addressing these ethical aspects, the project upheld the principles of responsible data usage and model development, ensuring that the findings were not only accurate but also ethically sound and reliable.

# Conclusion

These projects collectively represent my journey in mastering data science concepts and methodologies, embodying the program's learning objectives from leveraging technology in data handling to applying ethical considerations in predictive modeling. This collection of projects demonstrates not just a theoretical understanding, but also an applied understanding of how to use these data science principles, concepts, and methodologies in various contexts.

# Appendix

* [Project Portfolio Milestone Requirement Detailed Description and Procedures for Online students whose last term is Spring 2024 (Revised: November 2021)](https://github.com/bryancrigger/IST782-Applied-Data-Science-Portfolio/blob/main/MS%20ADS%20Portfolio%20Requirements%20-%20Spring%202024%20Term.pdf)
* [Applied Machine Learning: Evaluating U.S. Opioid Prescribers](https://github.com/bryancrigger/IST782-Applied-Data-Science-Portfolio/tree/main/IST707%20Applied%20Machine%20Learning)
* [Natural Language Processing: Classifying Spam vs Non-Spam Emails with NLP](https://github.com/bryancrigger/IST782-Applied-Data-Science-Portfolio/tree/main/IST664%20Natural%20Language%20Processing)
* [Big Data Analytics: Identifying most Important Factors in Calculating Home Price](https://github.com/bryancrigger/IST782-Applied-Data-Science-Portfolio/tree/main/IST718%20Big%20Data%20Analytics)
* [Deep Learning in Practice: Pistachio Image Classification](https://github.com/bryancrigger/IST782-Applied-Data-Science-Portfolio/tree/main/IST691%20Deep%20Learning%20in%20Practice)